Structural fingerprints of molecular evolution

Nature is using the same fold again and again to produce protein variants with comparable structures and adjusted biochemical properties.

The DNA molecule encodes all information required to create and sustain life. Mutations to DNA can dramatically alter the appearance, adaptation to the environment, and health of organisms. Mutations have produced the fantastically large number of species —estimated at five to sixty million— we see on earth. How many more variations can there be of DNA molecules? Are there meaningful constraints on the diversity of the genetic code?

Computational biologists like CS Professor Ron Elber are attempting to answer such fundamental biological questions using computational methods, starting at the smallest scale of biological activity. DNA codes proteins, which are the prime molecular machines of the cell. Proteins are linear polymers of amino acids. Since the amino-acid sequence determines all protein properties, including its three-dimensional shape, most evolutionary studies of proteins have focused on comparing sequences of amino acids.

Elber’s group is taking a different tack.

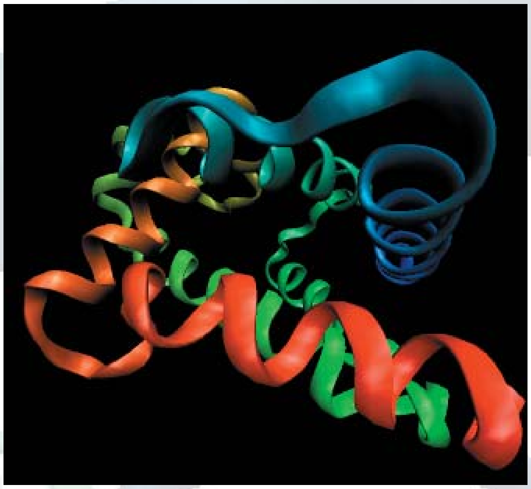
“Sequence comparison is effective in detecting closely related evolutionary changes, but it is not the best when remote evolutionary pairs are considered,” says Elber. “An interesting empirical observation is that protein structures are better preserved than their sequences. Nature is using the same fold again and again to produce protein variants with comparable structures and biochemical properties, so my group uses structural information for detecting remote evolutionary relationships.”

This idea of using protein structure rather than amino-acid sequences to uncover evolutionary patterns has led to some major breakthroughs.

In 2000, Steve Tanksley (Plant Breeding) and his co-workers found a gene that controls the size of the tomato fruit, but the evolutionary relationship of this gene with other known genes could not be identified based on sequence similarity. So they called Elber for help. Using his LOOPP software for matching protein sequences to shapes, Elber was able to determine within a few minutes that the tomato gene was remarkably similar to a human gene that controls cell division and growth (in fact, some human cancers result when this gene malfunctions).

“It is astounding that this kind of research could be done at the speed at which it was done,” says Tanksley. “This would have been impossible just a few years ago.”

Besides clarifying the molecular mechanisms that control tomato size, the study proposed an evolutionary pathway from the wild tomato to the domestic fruit.



The evolutionary capacity of about 4,000 proteins was determined. The capacity of the above protein (*Ascaris hemoglobin*) is 10^{190}.

Such successes on empirical data have led Elber and CS Professor Jon Kleinberg to ask a more theoretical question: how many distinct protein sequences can fold into a particular 3-dimensional shape?

This number, which they call the *evolutionary capacity of the protein*, can in principle be very large (since there are twenty types of amino acids, the sequence space of a protein of length L, L ≈ 100–1000, is 20^L), so it cannot be determined in reasonable time by direct enumeration.

Kleinberg, Elber, and their students Leonid Meyerguz and David Kempe have found that the problem is closely related to the well-known Knapsack problem, and this connection has led them to invent a fast randomized algorithm for estimating the evolutionary capacity of a protein. The figure on this page shows the protein *Ascaris hemoglobin*, which the

CS Professor Ron Elber: Sequence comparison is effective in detecting closely related evolutionary changes, but it is not the best when remote evolutionary pairs are considered. An interesting empirical observation is that protein structures are better preserved than their sequences.

team estimates has an evolutionary capacity of about 10^{190}.

"Using this algorithm, we have studied the evolutionary capacity of all known protein shapes," says Kleinberg. "The capacity of existing shapes is vastly larger than sequence spaces already explored by nature, suggesting that currently, it is not a limiting evolutionary factor."

Interestingly, it appears that proteins occurring in nature are not optimal from a structural point of view —Elber and Kleinberg found alternative sequences that led to similar but more stable structures. Have they improved on Mother Nature? Or are there additional factors that explain why Nature has chosen to construct proteins the way she has?

These kinds of fundamental questions about the nature of life on earth will keep computational biologists busy for many years.

Bridging the Rift: Promoting research and education and peace

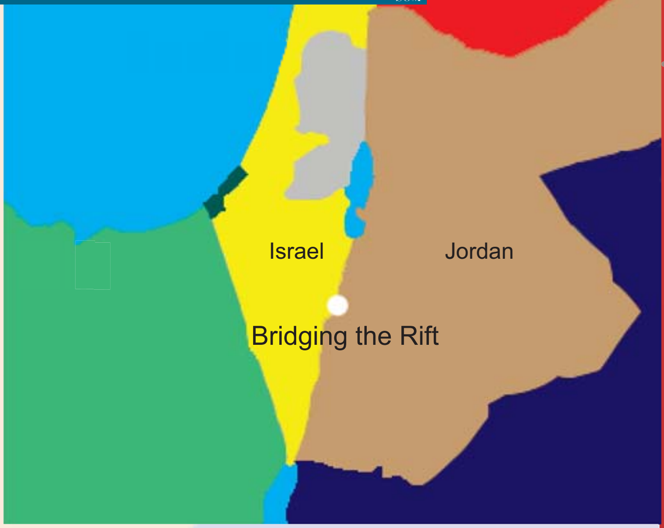
On 9 March 2005, the cornerstone of the *Bridging the Rift* Center (BTR) was laid in the desert, 43 miles south of the Dead Sea, between Israel and Jordan. The Cornell president and others, including CS's Bob Constable, took part, as did the Israeli and Jordanian ministers of education, the Israeli finance minister, the Jordanian minister of planning and international cooperation, and Mati Kochavi of the Bridging the Rift Foundation, which is providing the seed money.

BTR will be a life sciences research complex, created to educate grad students from both sides of the border, on 150 acres donated by Israel and Jordan. Israeli and Jordanian students will study side by side, along with grad students from Cornell and Stanford. Cornell and Stanford, substantial partners in this venture, will offer doctoral degrees at BTR, and their faculty will participate along with faculty from Israel and Jordan.

The main research project of BTR is the *Library of Life*. The goal is to assemble a digital catalog and living samples of all microbe, fungi, plants, insects, vertebrates, and invertebrates in the region, creating a Library of the Desert. Cornell professor of Plant Breeding Steven Tanksley conceived the idea. CS professor Ron Elber is the Director of the Library of Life.

The Library is expected to be a global research and education resource, but this will require the development of novel search, analysis, and modeling tools. Because of this, other CS faculty will be involved, including Rich Caruana, Johannes Gehrke, Dan Huttenlocher, Uri Keich, and Jayavel Shanmugasundaram.

BTR is becoming one of the most prominent and positive programs in the Middle East. King Abdullah II of Jordan described BTR as "bigger than Jordan and Israel", and Prime Minister Sharon of Israel identified it as being of "first rank strategic importance for the region". By spearheading this project, our department is not only doing excellent research but is contributing in a small way to peace in a troubled part of the world.



Juris Hartmanis receives the CRA Distinguished Service Award for his service in the areas of government affairs, professional societies, publications, conferences, and leadership, which had a major impact on computing research.

Dexter Kozen is the Class of 1960 Scholar, Williams College.

Joe Halpern is named the Milner Lecturer at the University of Edinburgh.

Greg Morrisett and students Steve Zdancewic and Dan Grossman receive the Best Paper Award in the European Association for Programming Languages and Systems Conference on Principles, Logics, and Implementation of High-Level Programming Languages.

Keshav Pingali and his students receive the Best paper Award at the International Conference of Supercomputing.

Former students John Belizaire and Julian Pelenur sell their company, Theory Center, Inc. The one-year-old company, a leading provider of Java Beans, was sold to BEA Systems for $100 million.

Bill Arms becomes the Series Editor of the MIT Press series on *Digital Libraries and Electronic Publishing*.

Johannes Gehrke publishes the second edition of *Database Management Systems* (McGraw Hill), with Ragu Ramakrishnan.

David Schwartz publishes *Introduction to UNIX* (Prentice Hall) and *Introduction to Maple* (Prentice Hall).

Under the leadership of Tom Coleman, the Cornell Theory Center opens the Financial Solutions Center on Broad Street in Manhattan.

Bart Selman's work on phase transitions and complexity is featured in *The New York Times*.

Gün Sirer, Golan Yona join.

2000

Ramin Zabih receives a joint appointment with the Cornell Medical School, the first such joint appointment at Cornell.

Jon Kleinberg receives the Best Paper Award, ACM Symposium on Principles of Database Systems.

Eva Tardos is elected to the American Academy of Arts & Sciences.

Juris Hartmanis receives the Lielo Medal from the Latvian Academy of Sciences. This highest award given by the Academy to scientists of Latvia and of foreign countries is for outstanding creative contributions.